

Pareto analysis based on records

M. Doostparast^{1,*} and N. Balakrishnan²

¹*Department of Statistics, School of Mathematical Sciences,
Ferdowsi University of Mashhad, P. O. Box 91775-1159, Mashhad, Iran*

²*Department of Mathematics and Statistics, McMaster University,
Hamilton, Ontario, Canada L8S 4K1*

Abstract

Estimation of the parameters of an exponential distribution based on record data has been treated by Samaniego and Whitaker (1986) and Doostparast (2009). Recently, Doostparast and Balakrishnan (2011) obtained optimal confidence intervals as well as uniformly most powerful tests for one- and two-sided hypotheses concerning location and scale parameters based on record data from a two-parameter exponential model. In this paper, we derive optimal statistical procedures including point and interval estimation as well as most powerful tests based on record data from a two-parameter Pareto model. For illustrative purpose, a data set on annual wages of a sample of production-line workers in a large industrial firm is analyzed using the proposed procedures.

Keywords and phrases: Generalized likelihood ratio test; Invariant test; Monotone likelihood ratio; Shortest-width confidence interval; Two-parameter Pareto model; Uniformly most powerful test.

1 Introduction

Let X_1, X_2, X_3, \dots be a sequence of continuous random variables. X_k is a lower record value if it is smaller than all preceding values X_1, X_2, \dots, X_{k-1} and by definition, X_1 is taken as the first lower record value. An analogous definition can be provided for upper record values. **Such data may be represented by $(\mathbf{R}, \mathbf{K}) := (R_1, K_1, \dots, R_m, K_m)$, where R_i is the i -th record value meaning new minimum (or maximum) and K_i is the number of trials following the observation of R_i that are needed to obtain a new record value R_{i+1} . Throughout this paper, we denote the observed value of these record data by $(\mathbf{r}, \mathbf{k}) := (r_1, k_1, \dots, r_m, k_m)$.** Record statistics arise naturally in

*Corresponding author.

E-mail addresses: doostparast@math.um.ac.ir (M. Doostparast), bala@mcmaster.ca (N. Balakrishnan).

many practical problems and in applied fields such as athletic events (Kuper and Sterken, 2003), Biology (Krug and Jain, 2005), catastrophic loss (Hsieh, 2004 and Pfeifer, 1997), climate research (Benestad, 2003), financial markets (Bradlow and Park, 2007 and de Haan *et al.*, 2009), industrial application (Samaniego and Whitaker, 1986 and 1988), spatial patterns (Yang and Lee, 2007), and traffic analysis (Glick, 1978). Hence, finding optimal statistical inferential procedures based on record data becomes very important and useful from a data-analysis point of view.

The rest of this article is organized as follows. In Section 2, we present briefly the notation to be used through out the paper and also the form of Pareto distribution to be studied here. In Section 3, we describe the basic form of record data to be considered and the corresponding likelihood function. In Section 4, we discuss the optimal point estimation of the Pareto parameters, while the interval estimation is handled in Section 5. Tests of hypotheses concerning the parameters are discussed in Section 6 and finally a numerical example is presented in Section 7 in order to illustrate all the inferential procedures developed here.

2 Some Preliminaries

Throughout the paper, we will use the following notation:

$Exp(\mu, \sigma)$: Exponential distribution with location μ and scale σ
$Gamma(n, \sigma)$: Gamma distribution with shape n and scale σ
$Par(\beta, \alpha)$: Pareto distribution with scale β and shape α
χ_v^2	: Chi-square distribution with v degrees of freedom
$\chi_{v,p}^2$: $100\gamma^{th}$ percentile of the chi-square distribution with v degrees of freedom
(\mathbf{r}, \mathbf{k})	: $(r_1, k_1, \dots, r_m, k_m)$
(\mathbf{R}, \mathbf{K})	: $(R_1, K_1, \dots, R_m, K_m)$
T_m	: $\sum_{i=1}^m K_i$, the time of occurrence of the m -th record
T_1^*	: $\sum_{i=1}^m K_i (\log R_i - \log \beta)$
T_2^*	: $\sum_{i=1}^{m-1} K_i (\log R_i - \log R_m)$
$\Gamma(r)$: $\int_0^\infty x^{r-1} e^{-x} dx$, the complete gamma function
$\hat{\theta}_M$: Maximum likelihood estimator of θ
$\hat{\theta}_U$: Unbiased estimator of θ

A random variable X is said to have a Pareto distribution, denoted by $X \sim Par(\beta, \alpha)$,

if its cumulative distribution function (cdf) is

$$F(x; \beta, \alpha) = 1 - \left(\frac{\beta}{x}\right)^\alpha, \quad x \geq \beta > 0, \alpha > 0, \quad (2.1)$$

and the probability density function (pdf) is

$$f(x; \beta, \alpha) = \alpha \beta^\alpha x^{-(\alpha+1)}, \quad x \geq \beta > 0, \alpha > 0. \quad (2.2)$$

For a through discussion on various properties and applications and different forms of Pareto distribution, one may refer to Arnold (1983) and Johnson, Kotz and Balakrishnan (1994).

3 Form of Data

As in Samaniego and Whitaker (1986) and Doostparast (2009), our starting point is a sequence of independent random variables X_1, X_2, X_3, \dots drawn from a fixed cdf $F(\cdot)$ and pdf $f(\cdot)$. We assume that only successive minima are observable, so that the data may be represented as $(\mathbf{r}, \mathbf{k}) := (r_1, k_1, r_2, k_2, \dots, r_m, k_m)$, where r_i is the value of the i -th observed minimum, and k_i is the number of trials required to obtain the next new minimum. The likelihood function associated with the sequence $\{r_1, k_1, \dots, r_m, k_m\}$ is given by

$$L(\mathbf{r}, \mathbf{k}) = \prod_{i=1}^m f(r_i) [1 - F(r_i)]^{k_i-1} I_{(-\infty, r_{i-1})}, \quad (3.1)$$

where $r_0 \equiv \infty$, $k_m \equiv 1$, and $I_A(x)$ is the indicator function of the set A .

The above described scheme is known as *inverse sampling scheme*. Under this scheme, items are presented sequentially and sampling is terminated when the m -th minimum is observed. In this case, the total number of items sampled is a random number, and K_m is defined to be one for convenience. There is yet another common scheme called *random sampling scheme* that is discussed in the literature. Under this scheme, a random sample Y_1, \dots, Y_n is examined sequentially and successive minimum values are recorded. In this setting, we have $N^{(n)}$, the number of records obtained, to be random and, given a value of m , we have in this case $\sum_{i=1}^m K_i = n$.

Remark Doostparast and Balakrishnan (2011) derived classical estimators for $Exp(\theta, \sigma)$ -model under both inverse and random sampling schemes, and also discussed associated cost-benefit analysis.

4 Point Estimation

Let us now assume that the sequence $\{R_1, K_1, \dots, R_m, K_m \equiv 1\}$ is arising from $Par(\beta, \alpha)$ in (2.1). Then, the likelihood function in (3.1) becomes

$$L(\beta, \alpha; \mathbf{r}, \mathbf{k}) = \frac{\alpha^m \beta^\alpha \sum_{i=1}^m k_i}{\prod_{i=1}^m r_i^{\alpha k_i + 1}}, \quad 0 < \beta \leq r_m, \quad \alpha > 0, \quad (4.1)$$

and so the log-likelihood function is

$$l(\beta, \alpha; \mathbf{r}, \mathbf{k}) = m \ln \alpha - \alpha \sum_{i=1}^m k_i (\ln r_i - \ln \beta) - \sum_{i=1}^m \ln r_i, \quad 0 < \beta \leq r_m, \quad \alpha > 0. \quad (4.2)$$

Since $\frac{\partial}{\partial \beta} l(\beta, \alpha; \mathbf{r}, \mathbf{k}) = \alpha \beta^{-1} \sum_{i=1}^m k_i > 0$, $l(\beta, \alpha; \mathbf{r}, \mathbf{k})$ is increasing with respect to β . This implies that

$$\hat{\beta}_M = R_m. \quad (4.3)$$

Substituting (4.3) in (4.2), the maximum likelihood estimate of α is readily obtained as

$$\hat{\alpha}_M = \frac{m}{T_2^\star}. \quad (4.4)$$

Furthermore, $(\sum_{i=1}^m K_i \ln R_i, T_m, R_m)$ is a joint sufficient statistic for (β, α) .

Corollary 4.1 *It can be shown that T_1^\star and T_2^\star are distributed as $\text{Gamma}(m, \alpha^{-1})$ and $\text{Gamma}(m-1, \alpha^{-1})$, respectively.*

From Corollary 4.1, an unbiased estimator for α is given by

$$\hat{\alpha}_U = \frac{m-1}{T_2^\star}.$$

In the following, we show that $\hat{\alpha}_M$ dominates $\hat{\alpha}_U$, under square error (SE) loss function. In other words, $\hat{\alpha}_U$ is inadmissible under SE loss function. First, we need the following lemma.

Lemma 4.2 *Suppose X has a $\text{Gamma}(\nu, \tau)$ distribution. Then,*

$$E(X^{-k}) = \tau^{-k} \frac{\Gamma(\nu - k)}{\Gamma(\nu)}, \quad k < \nu.$$

Proposition 4.3 *For $m > 4$, under the SE loss function, $\hat{\alpha}_M$ dominates $\hat{\alpha}_U$.*

Proof From Lemma 4.2, we have

$$\begin{aligned}
MSE(\hat{\alpha}_M) &:= E(\hat{\alpha}_M - \alpha)^2 \\
&= E\left(\frac{m}{T_2^*} - \alpha\right)^2 \\
&= m^2\alpha^2 \frac{\Gamma(m-1-2)}{\Gamma(m-1)} + \alpha^2 - 2\alpha^2 m \frac{\Gamma(m-1-1)}{\Gamma(m-1)} \\
&= \alpha^2 \left\{ \frac{m^2}{(m-2)(m-3)} + 1 - \frac{2m}{m-2} \right\} \\
&= \alpha^2 \frac{(m+6)}{(m-2)(m-3)}. \tag{4.5}
\end{aligned}$$

Replacing m with $m-1$ in (4.5), we immediately have

$$MSE(\hat{\alpha}_U) = \alpha^2 \frac{(m+5)}{(m-3)(m-4)}. \tag{4.6}$$

Thus, the efficiency of $\hat{\alpha}_M$ with respect to $\hat{\alpha}_U$ is given by

$$\begin{aligned}
EFF(\hat{\alpha}_M, \hat{\alpha}_U) &= \frac{MSE(\hat{\alpha}_U)}{MSE(\hat{\alpha}_M)} \\
&= \frac{(m+5)}{(m-3)(m-4)} \frac{(m-2)(m-3)}{(m+6)} \\
&= \frac{(m+5)(m-2)}{(m+6)(m-4)} \\
&= 1 + \frac{m+14}{(m+6)(m-4)} \\
&> 1, \tag{4.7}
\end{aligned}$$

which is the desired result. \square

One can easily check that the bias of $\hat{\alpha}_M$, under the SE loss function, is

$$B_{SE}(\hat{\alpha}_M, \alpha) = E(\hat{\alpha}_M) - \alpha = 2\alpha/(m-2).$$

It may be noted that

$$\lim_{m \rightarrow \infty} EFF(\hat{\alpha}_M, \hat{\alpha}_U) = 1.$$

This is to be expected since $\hat{\alpha}_M$ and $\hat{\alpha}_U$ are equivalent for large values of m . Figure 1 shows the relative efficiency of $\hat{\alpha}_M$ with respect to $\hat{\alpha}_U$.

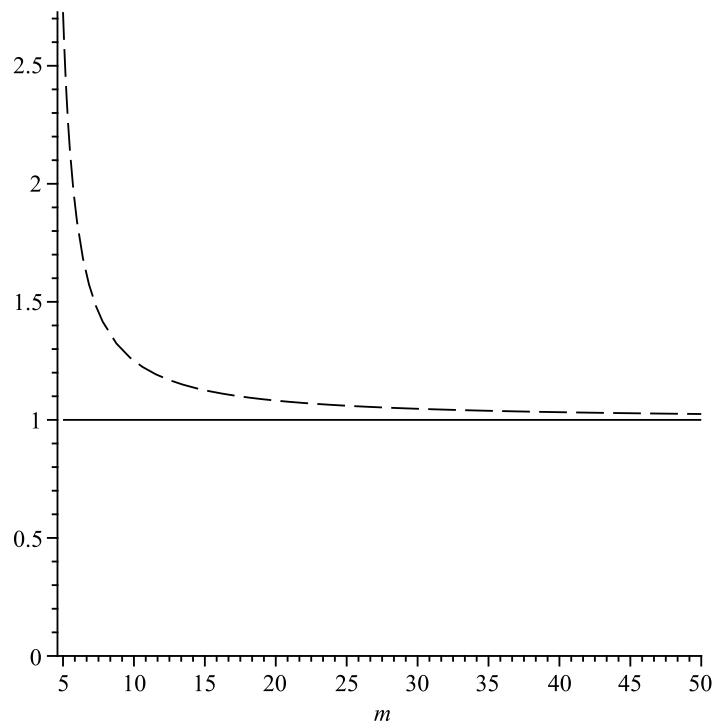


Figure 1: Relative efficiency of $\hat{\alpha}_M$ with respect to $\hat{\alpha}_U$ given by (4.7).

5 Confidence intervals

Suppose we observe (\mathbf{r}, \mathbf{k}) from a two-parameter Pareto distribution in (2.1). In this section, we discuss the construction of exact confidence intervals for the two parameters in different cases.

5.1 α known

Suppose the shape parameter α is known. Then, from (4.1), we have (T_m, R_m) to be a joint sufficient statistic for β . Since T_m is distributed free from parent distribution (Glick, 1978), we consider two approaches for obtaining confidence intervals for β on the basis of record data.

Unconditional method

To obtain a confidence interval for β , we need the following lemma due to Doostparast and Balakrishnan (2011).

Lemma 5.1 *Suppose we observe (\mathbf{r}, \mathbf{k}) from a two-parameter $Exp(\mu, \sigma)$ -distribution. Then*

$$\frac{R_m - \mu}{\sigma} \sim g(x; m) = \frac{\{-\ln(1 - \exp(-x))\}^{m-1}}{\Gamma(m)} \exp(-x), \quad x \geq 0. \quad (5.1)$$

Lemma 5.2 *If $X \sim Par(\beta, \alpha)$, then $\ln X \sim Exp(\ln \beta, \alpha^{-1})$.*

From Lemmas 5.1 and 5.2, it can be shown in this case that

$$\alpha (\ln R_m - \ln \beta) \sim g(x; m). \quad (5.2)$$

This implies that $-\ln(1 - \exp\{-\alpha (\ln R_m - \ln \beta)\})$ has a gamma distribution with parameters $(m, 1)$, and therefore,

$$-2 \ln \left(1 - \left(\frac{\beta}{R_m} \right)^\alpha \right) \sim \chi_{(2m)}^2. \quad (5.3)$$

Hence, an equi-tailed $100(1 - \gamma)\%$ confidence interval for β is given by

$$I_{ET,1}(\beta) = \left(R_m \sqrt[m]{1 - \exp \left\{ -\frac{\chi_{2m, \frac{\gamma}{2}}^2}{2} \right\}}, R_m \sqrt[m]{1 - \exp \left\{ -\frac{\chi_{2m, 1 - \frac{\gamma}{2}}^2}{2} \right\}} \right). \quad (5.4)$$

Suppose we restrict our attention to intervals of the form (aR_m, bR_m) , where $0 < a < b$. Since the function $g(x; m)$ in (5.1) is decreasing with respect to x for every $m \geq 1$, the $100(1 - \gamma)\%$ confidence interval for β with minimum width in this subclass of intervals is

$$I_{ML,1}(\beta) = \left(R_m \exp \left\{ -\frac{g_{m, 1-\gamma}}{\alpha} \right\}, R_m \right), \quad (5.5)$$

where $g_{m, \gamma}$ is 100γ -th percentile of the pdf $g(x; m)$ in (5.1).

Conditional method

Since minimum of a random sample of size n from $Exp(0, \sigma)$ has a $Exp(0, \sigma/n)$ -distribution [see Arnold, Balakrishnan and Nagaraja (1992)], conditional on $T_m = j$ for $j \geq m$, the random variable $\alpha j(\ln R_m - \ln \beta)$ has a standard exponential distribution. Therefore, a conditional equi-tailed $100(1 - \gamma)\%$ confidence interval for β is given by

$$\left(R_m \sqrt[\alpha j]{\frac{\gamma}{2}}, R_m \sqrt[\alpha j]{1 - \frac{\gamma}{2}} \right).$$

This implies that an equi-tailed $100(1 - \gamma)\%$ confidence interval for β is

$$I_{ET,1,C}(\beta) = \left(R_m \left(\frac{\gamma}{2} \right)^{\frac{1}{\alpha T_m}}, R_m \left(1 - \frac{\gamma}{2} \right)^{\frac{1}{\alpha T_m}} \right). \quad (5.6)$$

The expected width of the interval in (5.4) is

$$L(I_{ET,1}(\beta)) = E(R_m) \left(\sqrt[\alpha]{1 - \exp \left\{ -\frac{\chi_{2m,1-\frac{\gamma}{2}}^2}{2} \right\}} - \sqrt[\alpha]{1 - \exp \left\{ -\frac{\chi_{2m,\frac{\gamma}{2}}^2}{2} \right\}} \right), \quad (5.7)$$

while the expected width of the interval in (5.6) is

$$\begin{aligned} L(I_{ET,C,1}(\beta)) &= E \left(R_m \left\{ \left(1 - \frac{\gamma}{2} \right)^{\frac{1}{\alpha T_m}} - \left(\frac{\gamma}{2} \right)^{\frac{1}{\alpha T_m}} \right\} \right) \\ &= \sum_{j=m}^{\infty} E \left(R_m \left\{ \left(1 - \frac{\gamma}{2} \right)^{\frac{1}{\alpha T_m}} - \left(\frac{\gamma}{2} \right)^{\frac{1}{\alpha T_m}} \right\} | T_m = j \right) P(T_m = j) \\ &= \sum_{j=m}^{\infty} \left\{ \left(1 - \frac{\gamma}{2} \right)^{\frac{1}{\alpha j}} - \left(\frac{\gamma}{2} \right)^{\frac{1}{\alpha j}} \right\} E(R_m | T_m = j) P(T_m = j). \end{aligned}$$

Again, since the minimum of a random sample of size n from $Exp(0, \sigma)$ has a $Exp(0, \sigma/n)$ -distribution, from Lemma 5.2, we conclude that $\ln R_m | (T_m = j)$ has a $Exp(\ln \beta, (\alpha j)^{-1})$ distribution. So,

$$\begin{aligned} E(R_m | T_m = j) &= E(\exp\{\ln R_m\} | T_m = j) \\ &= \int_{\ln \beta}^{\infty} e^y j \alpha e^{-j \alpha (y - \ln \beta)} dy \\ &= \frac{\beta j \alpha}{j \alpha - 1}. \end{aligned}$$

Therefore,

$$\begin{aligned}
L(I_{ET,C,1}(\beta)) &= \sum_{j=m}^{\infty} \left\{ \left(1 - \frac{\gamma}{2}\right)^{\frac{1}{\alpha j}} - \left(\frac{\gamma}{2}\right)^{\frac{1}{\alpha j}} \right\} \frac{\beta j \alpha}{j \alpha - 1} P(T_m = j) \\
&= E \left(\left\{ \left(1 - \frac{\gamma}{2}\right)^{\frac{1}{\alpha T_m}} - \left(\frac{\gamma}{2}\right)^{\frac{1}{\alpha T_m}} \right\} \frac{\beta T_m \alpha}{T_m \alpha - 1} \right) \\
&= \beta E \left(\left\{ \left(1 - \frac{\gamma}{2}\right)^{\frac{1}{\alpha T_m}} - \left(\frac{\gamma}{2}\right)^{\frac{1}{\alpha T_m}} \right\} \frac{T_m \alpha}{T_m \alpha - 1} \right).
\end{aligned}$$

Hence, for computing the expected width of the interval in (5.6), we need the probability mass function of T_m . From Sibuya and Nishimura (1997), we have

$$P(T_m = j) = \frac{1}{j!} \left[\begin{matrix} j-1 \\ m-1 \end{matrix} \right], \quad j \geq m, \quad (5.8)$$

where brackets $[\]$ denote unsigned Stirling numbers of the first kind defined by the polynomial identity

$$z^{[n]} := z(z+1) \cdots (z+n-1) = \sum_{m=1}^n \left[\begin{matrix} n \\ m \end{matrix} \right] z^m.$$

Now, let $H(\cdot)$ be an arbitrary function. Then, Doostparast and Balakrishnan (2009) showed that

$$E(H(T_m)) = E \left(\frac{T_m H(T_m - 1)}{T_m - 2} \right) - E \left(\frac{H(T_{m-1})}{T_{m-1} - 1} \right) \quad (5.9)$$

and this formula may be used for obtaining the required expectations by taking a suitable choice for the function $H(\cdot)$. However, no explicit expression seems possible for $E \left(a^{\frac{1}{\alpha T_m}} \frac{T_m \alpha}{T_m \alpha - 1} \right)$ and so a simulation study was carried out to generate sequences of independent observations based on which the desired estimates were calculated for $E \left(a^{\frac{1}{\alpha T_m}} \frac{T_m \alpha}{T_m \alpha - 1} \right)$ in the illustrative examples.

Remark The theory of uniformly most powerful (UMP) one-sided test can be applied to the problem of obtaining a lower or upper bounds. In Section 6, we will obtain uniformly most accurate (UMA) lower and upper bounds for β .

5.2 β known

If β is known, then T_1^* is a complete sufficient statistic for α , and so confidence intervals can be based on this statistic. Since T_1^* is distributed as $Gamma(m, \alpha^{-1})$, we have

$$2\alpha T_1^* \sim \chi_{(2m)}^2. \quad (5.10)$$

γ	m					
	2	3	4	5	6	7
0.10	0.167630	0.882654	1.874590	3.017327	4.258219	5.569586
	7.864292	10.958349	13.892227	16.710795	19.446252	22.118958
0.05	0.084727	0.607001	1.425002	2.413920	3.516159	4.700465
	9.530336	12.802444	15.896592	18.860434	21.728898	24.524694
0.01	0.017469	0.263963	0.785646	1.497847	2.344412	3.291176
	13.285448	16.901320	20.295553	23.532765	26.653130	29.683220

Table 1: Values of a (the upper figure) and b (the lower figure) in (5.12) for $\gamma = 0.01, 0.05, 0.1$ and different choices of m .

Therefore, in practice, one may use equi-tailed $100(1 - \gamma)\%$ interval of the form

$$I_{ET,1}(\alpha) := \left(\frac{\chi_{2m,\gamma/2}^2}{2T_1^*}, \frac{\chi_{2m,1-(\gamma/2)}^2}{2T_1^*} \right).$$

Suppose we restrict ourselves to a class of intervals of the form

$$I_1(a, b) = \left(\frac{a}{2T_1^*}, \frac{b}{2T_1^*} \right), \quad 0 < a < b. \quad (5.11)$$

We then need to find a and b that minimizes the width of the interval in (5.11) subject to the confidence coefficient being $1 - \gamma$. Using Lagrange method, we then need to solve the following equations for a and b , determining $I_{ML}(\alpha)$, as

$$\int_a^b h_{2m}(x)dx = 1 - \gamma \quad \text{and} \quad h_{2m}(a) = h_{2m}(b), \quad (5.12)$$

where $h_v(x)$ is the density function of a chi-square distribution with v degrees of freedom given by

$$h_v(x) = \frac{1}{2^{v/2}\Gamma(\frac{v}{2})} x^{v/2-1} \exp\left(-\frac{x}{2}\right), \quad x > 0. \quad (5.13)$$

Table 1 presents values of a and b up to six decimal places that satisfy the conditions in (5.12).

Suppose that the random variable X has a $Gamma(v, \tau)$ -distribution, where v is a known constant. A UMP test does not exist for testing $H_0 : \tau = \tau_0$ against the alternative $H_1 : \tau \neq \tau_0$ (Lehmann, 2000, p. 111). So, there are no UMA bounds for α . However, the acceptance region of the UMP unbiased test is

$$C_1 \leq \frac{2X}{\tau_0} \leq C_2,$$

where C_1 and C_2 are obtained from the equations

$$\int_{C_1}^{C_2} h_{2v}(x)dx = 1 - \gamma \quad \text{and} \quad C_1^v e^{-C_1/2} = C_2^v e^{-C_2/2}.$$

This yields UMA unbiased bounds for α as

$$\left(\frac{C_1}{2T_1^*}, \frac{C_2}{2T_1^*} \right), \tag{5.14}$$

with

$$\int_{C_1}^{C_2} h_{2m}(x)dx = 1 - \gamma \quad \text{and} \quad C_1^m e^{-C_1/2} = C_2^m e^{-C_2/2}.$$

Corollary 5.3 *UMA unbiased and minimum width intervals in the class (5.11) given by (5.14) and (5.12), respectively, are identical.*

Remark From Lehmann (2005, p. 72, Theorem 3.5.1) and (5.10), the acceptance region of the most powerful test of $H_0 : \alpha = \alpha_0$ against $H_1 : \alpha < \alpha_0$ is $2\alpha_0 T_1^* \leq C_u$, where C_u is determined by the equation

$$\int_0^{C_u} h_{2m}(x)dx = 1 - \gamma.$$

Therefore, $\frac{\chi_{2m,1-\gamma}^2}{2T_1^*}$ is a UMA upper confidence bound for α . Similarly, $\frac{C_L}{2T_1^*}$ is a UMA lower confidence bound for α , where C_L is such that

$$\int_{C_L}^{\infty} h_{2m}(x)dx = 1 - \gamma,$$

or

$$\int_0^{C_L} h_{2m}(x)dx = \gamma.$$

That is, $\frac{\chi_{2m,\gamma}^2}{2T_1^*}$ is a uniformly most accurate lower confidence bound for α (without the restriction of unbiasedness). For more details, one may refer to Lehmann (2005) and Pachares (1961) for tables of C_1 and C_2 .

5.3 β and α both unknown

From (4.1), the statistic (T_2^*, T_m, R_m) is jointly sufficient for β and α . Therefore, confidence intervals may be developed based on this statistic.

Confidence interval for α

From Lemma 5.2, α^{-1} is a scale parameter for data $(R'_1, K_1, \dots, R'_m, K_m)$, where $R'_i = \ln R_i$ for $1 \leq i \leq m$. Therefore, the assumption that the limits remain unchanged upon the addition of a constant to all log-record values (R'_i) seems reasonable and leads to intervals depending only on T_2^* . For convenience, we restrict ourselves to multiples of T_2^* for intervals of the form

$$I_2(a, b) = \left(\frac{a}{2T_2^*}, \frac{b}{2T_2^*} \right), \quad 0 < a < b. \quad (5.15)$$

Now, since T_2^* is distributed as $\text{Gamma}(m-1, \alpha^{-1})$, we have $2\alpha T_2^* \sim \chi_{2(m-1)}^2$. So, we can use the conditions in (5.12) for obtaining the minimum width confidence interval simply by replacing m and T_1^* by $m-1$ and T_2^* , respectively.

Confidence interval for β

First, we need the following lemma of Doostparast and Balakrishnan (2011).

Lemma 5.4 *Suppose the random variable U is distributed with pdf $g(x; m)$ as in (5.1) ($m > 1$) and T is a chi-square random variable with ν degrees of freedom. If U and T are independent, then the pdf of $W := \frac{2U}{T}$ is given by*

$$f_W(w; m, \nu) = \frac{1}{w^{1+\nu/2} \Gamma(m) \Gamma(\nu/2)^2} \int_0^\infty \{ -\ln(1 - e^{-x}) \}^{m-1} x^{\nu/2} e^{-x(1+\frac{1}{w})} dx. \quad (5.16)$$

Since the random variable R_m and T_2^* are independent (Arnold *et al.*, 1998) and that $2\alpha T_2^*$ is distributed as chi-square with $2(m-1)$ degrees of freedom, by Lemma 5.4, we can conclude that

$$\frac{\ln R_m - \ln \beta}{T_2^*} \sim f_W(w; m, 2m-2). \quad (5.17)$$

So, an equi-tailed $100(1 - \gamma)\%$ confidence interval is given by

$$\left(R_m \left[\prod_{i=1}^m \left(\frac{R_m}{R_i} \right)^{K_i} \right]^{w_{1-\gamma/2}(m, 2m-2)}, R_m \left[\prod_{i=1}^m \left(\frac{R_m}{R_i} \right)^{K_i} \right]^{w_{\gamma/2}(m, 2m-2)} \right), \quad (5.18)$$

where $w_\gamma(m, \nu)$ is the 100γ -th percentile of the density in (5.16). For some choices of m and γ , the values of $w_\gamma(m, 2m-2)$ were obtained by Doostparast and Balakrishnan (2009) and these are presented in Table 2.

Restricting to intervals of the form (aR_m, bR_m) , where $0 < a < b$, the values of a and b which minimize the width in this subclass of intervals subject to the confidence coefficient

m	0.01	0.025	0.05	0.95	0.975	0.99
2	0.00068871	0.00200001	0.00462857	1.28171529	1.99110929	3.40093274
3	0.00005829	0.00018945	0.00048138	0.19893804	0.28399815	0.42122379
4	0.00000748	0.00002678	0.00007396	0.05887585	0.08503410	0.12545761
5	0.00000118	0.00000457	0.00001356	0.02116642	0.03157779	0.04776232
6	0.00000021	0.00000088	0.00000279	0.00829748	0.01289715	0.02026765
7	0.00000004	0.00000018	0.00000061	0.00339234	0.00551395	0.00908049
8	0.00000001	0.00000004	0.00000014	0.00141525	0.00240988	0.00415617

Table 2: Percentiles of the density in (5.16) for $m = 2, \dots, 8$ and $\nu = 2m - 2$.

being $1 - \gamma$ can be obtained by solving the following equations:

$$\begin{cases} f_W\left(\frac{a}{T_2^*}; m, 2m - 2\right) = f_W\left(\frac{b}{T_2^*}; m, 2m - 2\right), \\ \int_{a/T_2^*}^{b/T_2^*} f_W(x; m, 2m - 2) dx = 1 - \gamma, \end{cases} \quad (5.19)$$

where $f_W(w; m, \nu)$ is as in (5.16).

6 Tests of Hypotheses

In this section, we treat tests of hypotheses concerning the two parameters of the Pareto distribution in (2.1). To this end, we consider the following three cases.

6.1 α known

If α is known, then (T_m, R_m) is a joint sufficient statistic for β . Since T_m is an ancillary statistic (Glick, 1978), R_m is a partially sufficient statistic for β . The joint pdf of (\mathbf{R}, \mathbf{K}) given by (4.1) possesses the MLR property in R_m . From Theorem 2 of Lehmann (1997, p. 78) and (5.3), the UMP test of size γ for testing $H_0 : \beta \leq \beta_0$ against the alternative $H_1 : \beta > \beta_0$ is

$$\phi(\mathbf{r}, \mathbf{k}) = \begin{cases} 1, & R_m \geq \beta_0 (1 - \exp\{-\frac{1}{2}\chi_{2m, \gamma}^2\})^{-1/\alpha}, \\ 0, & R_m < \beta_0 (1 - \exp\{-\frac{1}{2}\chi_{2m, \gamma}^2\})^{-1/\alpha}. \end{cases} \quad (6.1)$$

By interchanging inequalities throughout, one obtains in an obvious way the solution for the dual problem. Thus, the UMP test of size γ for testing $H_0 : \beta \geq \beta_0$ against the alternative $H_1 : \beta < \beta_0$ is

$$\phi(\mathbf{r}, \mathbf{k}) = \begin{cases} 1, & R_m \leq \beta_0 (1 - \exp\{-\frac{1}{2}\chi_{2m, 1-\gamma}^2\})^{-1/\alpha}, \\ 0, & R_m > \beta_0 (1 - \exp\{-\frac{1}{2}\chi_{2m, 1-\gamma}^2\})^{-1/\alpha}. \end{cases} \quad (6.2)$$

The UMP tests in (6.1) and (6.2) imply that the UMP test for testing $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$ does not exist.

From (4.1), the likelihood ratio statistic is obtained as $\Lambda = (\beta_0/R_m)^{\alpha T_m}$ for $r_m \geq \beta_0$ and $\Lambda = 0$ for $r_m < \beta_0$. Thus, the critical region of the GLR test of level γ is $C = \{(\mathbf{r}, \mathbf{k}) : \alpha T_m(\log R_m - \log \beta_0) < c, \text{ or } R_m < \beta_0\}$. Since $\alpha T_m(\log R_m - \log \beta_0) | (T_m = j)$ has a standard exponential distribution and T_m is distributed free from the parent distribution (Glick, 1978), we have the following proposition.

Proposition 6.1 *Critical region of the GLR test of level γ for testing $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$ is given by*

$$C = \left\{ (\mathbf{r}, \mathbf{k}) : \left(\frac{\beta_0}{R_m} \right)^{\alpha T_m} < \gamma \text{ or } R_m < \beta_0 \right\}. \quad (6.3)$$

As mentioned earlier, the theory of UMP one-sided test can be applied to the problem of obtaining a lower or upper bound. Thus, from (6.1) and (6.2), $100(1 - \gamma)\%$ UMA lower and upper bounds for β are given by

$$\left(R_m \left(1 - \exp \left\{ -\frac{1}{2} \chi_{2m, \gamma}^2 \right\} \right)^{1/\alpha}, \infty \right)$$

and

$$\left(0, R_m \left(1 - \exp \left\{ -\frac{1}{2} \chi_{2m, 1-\gamma}^2 \right\} \right)^{1/\alpha} \right),$$

respectively.

6.2 β known

If β is known, the statistic T_1^* is a complete sufficient statistic, and so all inference can be based on it. Since T_1^* has a $Gamma(m, \alpha^{-1})$ -distribution, and has MLR in $-T_1^*$, the UMP tests of size γ for testing $H_0 : \alpha \leq \alpha_0$ against the alternative $H_1 : \alpha > \alpha_0$ and $H_0 : \alpha \geq \alpha_0$ against the alternative $H_1 : \alpha < \alpha_0$ are

$$\phi(\mathbf{r}, \mathbf{k}) = \begin{cases} 1, & 2\alpha_0 T_1^* \leq \chi_{2m, \gamma}^2, \\ 0, & \text{otherwise,} \end{cases} \quad (6.4)$$

and

$$\phi(\mathbf{r}, \mathbf{k}) = \begin{cases} 1, & 2\alpha_0 T_1^* \geq \chi_{2m, 1-\gamma}^2, \\ 0, & \text{otherwise,} \end{cases} \quad (6.5)$$

	m				
γ	1	2	3	4	5
0.01	0.0169	0.0589	0.3139	2.2636	20.7899
0.02	0.0332	0.1136	0.6004	4.3280	39.7496
0.03	0.0490	0.1658	0.8766	6.2658	57.6354
0.04	0.0647	0.2175	1.1370	8.1415	74.3884
0.05	0.0797	0.2664	1.3850	9.9380	90.4894
0.1	0.1517	0.4918	2.5377	18.0175	163.4582

Table 3: Quantiles of $X^m \exp\{-X/2\}$, where $X \sim \chi_{2m}^2$, for some choices of m and γ .

respectively. Therefore, the UMP test for testing $H_0 : \alpha = \alpha_0$ against the alternative $H_1 : \alpha \neq \alpha_0$ does not exist. One can easily show that the critical region of the GLR test of level γ for testing $H_0 : \alpha = \alpha_0$ against the alternative $H_1 : \alpha \neq \alpha_0$ is

$$C = \left\{ (\mathbf{r}, \mathbf{k}) : Z_1^m \exp \left\{ -\frac{1}{2} Z_1 \right\} < c^* \right\}, \quad (6.6)$$

where under H_0 , $Z_1 := 2\alpha_0 T_1^* \sim \chi_{2m}^2$ and c^* is chosen such that

$$\gamma = P_{\alpha=\alpha_0} \left(Z_1^m \exp \left\{ -\frac{1}{2} Z_1 \right\} < c^* \right).$$

Table 3 presents simulated critical values for applying the GLR test, obtained by Doostparast and Balakrishnan (2011), for some choices of m and γ .

6.3 Unknown β and α

Hypotheses tests for α

There is no UMP test for one-sided hypotheses on the scale parameter β . So, we restrict our attention to smaller classes of tests and seek UMP tests in these subclasses.

The family of densities $\{Par(\beta, \alpha) : \beta > 0, \alpha > 0\}$ remains invariant under translations $R'_i = R_i/c$, $0 < c < \infty$. Moreover, the hypotheses-testing problem remains invariant under the group of translations, that is, both families of pdfs $\{Par(\beta, \alpha), \alpha \geq \alpha_0\}$ and $\{Par(\beta, \alpha), \alpha < \alpha_0\}$ remain invariant. On the other hand, the joint sufficient statistic is (R_m, T_2^*, T_m) , which is transformed to $(R_m/c, T_2^*, T_m)$. It follows that the class of invariant tests consists of tests that are functions of T_2^* . Since $T_2^* \sim Gamma(m-1, \alpha^{-1})$, the pdf of T_2^* possesses the MLR property in $-T_2^*$, and it therefore follows that a UMP test rejects $H_0 : \alpha \leq \alpha_0$ if $T_2^* < c$, where c is determined from the size restriction. Hence, we have the following proposition which presents a UMP invariant test for one-sided hypotheses.

Proposition 6.2 *To test $H_0 : \alpha \leq \alpha_0$ against $H_1 : \alpha > \alpha_0$, a UMP invariant test of size γ is*

$$\phi(\mathbf{r}, \mathbf{k}) = \begin{cases} 1, & 2\alpha_0 T_2^* \leq \chi_{2m-2, \gamma}^2, \\ 0, & \text{otherwise,} \end{cases} \quad (6.7)$$

and to test $H_0 : \alpha \geq \alpha_0$ against $H_1 : \alpha < \alpha_0$, a UMP invariant test of size γ is

$$\phi(\mathbf{r}, \mathbf{k}) = \begin{cases} 1, & 2\alpha_0 T_2^* \geq \chi_{2m-2, 1-\gamma}^2, \\ 0, & \text{otherwise.} \end{cases} \quad (6.8)$$

There is no UMP test for testing $H_0 : \alpha = \alpha_0$ against the alternative $H_1 : \alpha \neq \alpha_0$. We therefore use the GLR procedure for this testing problem. The likelihood ratio function is given by

$$\Lambda = \left(\frac{Z_2}{2m} \right)^m \exp\{-m(Z_2/2m - 1)\}, \quad (6.9)$$

where $Z_2 = 2\alpha_0 T_2^*$ which, under H_0 , is distributed as chi-square with $2(m-1)$ degrees of freedom. Hence, the critical region of GLR test at level γ is

$$C = \{(\mathbf{r}, \mathbf{k}) : y^m \exp(-y/2) < a\}, \quad (6.10)$$

where a is chosen such that $\gamma = P(Z_2^m \exp\{-Z_2/2\} < a)$ and $Z_2 \sim \chi_{2(m-1)}^2$. Table 3 presents critical values for applying the GLR test for some choices of m and γ .

Hypotheses tests for β

In the case of unknown α , finding a UMP test for one- and two-sided hypotheses remains as an open problem. However, $\hat{\alpha}_{M,0} = \frac{m}{\sum_{i=1}^m K_i(\log R_i - \log \beta_0)}$ is the maximum likelihood estimator of α under $H_0 : \beta = \beta_0$. This fact and (4.1) yield the likelihood ratio statistic, for testing $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$, as

$$\Lambda = \begin{cases} \left(\frac{T_2^*}{T_0^*} \right)^m & \text{for } r_m \geq \beta_0 \\ 0 & \text{for } r_m < \beta_0 \end{cases}, \quad (6.11)$$

where $T_0^* = \sum_{i=1}^m K_i(\log R_i - \log \beta_0)$. But,

$$\frac{T_2^*}{T_0^*} = 1 - \frac{T_m(\log R_m - \log \beta_0)}{m\hat{\alpha}_{M,0}}.$$

Therefore, the critical region of the GLR test of level γ for testing $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$ is given by

$$C = \{(\mathbf{r}, \mathbf{k}) : T_m(\log R_m - \log \beta_0) > \hat{\alpha}_{M,0} C^* \text{ or } R_m < \beta_0\}, \quad (6.12)$$

Table 4: Annual wage data (in multiplies of 100 U.S. dollars).

112	154	119	108	112	156	123	103	115	107
125	119	128	132	107	151	103	104	116	140
108	105	158	104	119	111	101	157	112	115

Table 5: Record data arising from annual wage data with $m = 3$.

i	1	2	3
R_i	112	108	103
K_i	3	4	1

where C^* is obtained from the size restriction

$$\gamma = P_{\beta_0} (T_m(\log R_m - \log \beta_0) > \hat{\alpha}_{M,0} C^*) . \quad (6.13)$$

An explicit closed-form expression for C^* in (6.12) does not seem to be possible. But, one can use the following expression for computational purposes:

$$\begin{aligned} \gamma &= P_{\beta_0} (T_m(\log R_m - \log \beta_0) > \hat{\alpha}_{M,0} C^*) \\ &= \sum_{j=m}^{\infty} P_{\beta_0} (T_m(\log R_m - \log \beta_0) > \hat{\alpha}_{M,0} C^* | T_m = j) P(T_m = j) \\ &= \sum_{j=m}^{\infty} P_{\beta_0} \left(\frac{\log R_m - \log \beta_0}{\sum_{i=1}^m K_i (\log R_i - \log \beta_0)} > \frac{C^*}{mj} | T_m = j \right) \frac{1}{j!} \left[\begin{matrix} j-1 \\ m-1 \end{matrix} \right] . \end{aligned} \quad (6.14)$$

7 Numerical Example

Dyer (1981) reported annual wage data (in multiplies of 100 U.S. dollars) of a random sample of 30 production-line workers in a large industrial firm, as presented in Table 4. He determined that Pareto distribution provided an adequate fit for data. Assuming inverse sampling scheme with $m = 3$, the corresponding record data are presented in Table 5. From (4.3) and (4.4), the MLE of β and α on the basis of record data are obtained to be $\hat{\beta}_M = 103$ and $\hat{\alpha}_M = 6.804$, respectively. From (5.18), an equi-tailed 95% confidence interval for β is obtained to be

$$(90.877, 102.991) .$$

Similarly, a minimum-width 95% confidence interval for α in the class (5.15) is obtained as

$$(0.096, 10.807) .$$

For testing $H_0 : \alpha = 6$ against the alternative $H_1 : \alpha \neq 6$, we have $Z_2 = 2\alpha_0 T_2^* = 5.291$, which gives

$$Z_2^m \exp \left\{ -\frac{Z_2}{2} \right\} = 10.512.$$

From Table 3, we have $a = 0.2664$. Therefore, (6.10) implies that H_0 is not rejected. Since we can not find C^* in (6.12), we can not test the hypothesis $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$. In this case, one may conduct a simulation study and calculate the percentile of Λ in (6.11) by specifying β_0 , and then carry out a likelihood-ratio test.

Acknowledgements

The authors are grateful to anonymous referees and the associate editor for their useful suggestions and comments on an earlier version of this manuscript, which resulted in this improved version.

Bibliography

1. Arnold, B. C. (1983). *Pareto Distributions*, International Co-operative Publishing House, Fairland, MD.
2. Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics*, John Wiley & Sons, New York.
3. Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1998). *Records*, John Wiley & Sons, New York.
4. Bardlow, E. T. and Park, Y. (2007). Bayesian estimation of bid sequences in internet auction using a generalized record-breaking model, *Marketing Science*, **26**, 218–229.
5. Benestad, R. E. (2003). How often can we expect a record event?, *Climate Research*, **25**, 3–13.
6. de Haan, L., de Vries, C. G. and Zhou, C. (2009). The expected payoff to Internet auctions, *Extremes*, **12**, 219–238.
7. Doostparast, M. (2009). A note on estimation based on record data, *Metrika*, **69**, 69–80.
8. Doostparast, M. and Balakrishnan, N. (2011). Optimal record-based statistical procedures for the two-parameter exponential distribution, *Journal of Statistical Computation and Simulation*, DOI: 10.1080/00949655.2010.513979.

9. Dyer, D. (1981). Structural probability bounds for the strong Pareto laws, *The Canadian Journal of Statistics*, **9**, 71–77.
10. Glick, N. (1978). Breaking records and breaking boards, *American Mathematical Monthly*, **85**, 2–26.
11. Hsieh, P. (2004). A data-analytic method for forecasting next record catastrophe loss, *Journal of Risk and Insurance*, **71**, 309–322.
12. Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distribution- Vol. 1*, Second edition, John Wiley & Sons, New York.
13. Krug, J. and Jain, K. (2005). Breaking records in the evolutionary race, *Physica A*, **53**, 1–9.
14. Kuper, G. H. and Sterken, E. (2003). Endurance in speed skating: The development of world records, *European Journal of Operational Research*, **148**, 293–301.
15. Lehmann, E. L. (1997). *Testing Statistical Hypotheses*, Second edition, Springer-Verlag, New York.
16. Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, Third edition, Springer-Verlag, New York.
17. Pachares, J. (1961). Tables for unbiased tests on the variance of a normal population, *Annals of Mathematical Statistics*, **32**, 84–87.
18. Pfeifer D. (1997). A statistical model to analyse natural catastrophe claims by mean of record values, In: *Proceedings of the 28 International ASTIN Colloquium, Cairns, Australia*, August 10-12, 1997, The Institute of Actuaries of Australia.
19. Samaniego, F. J. and Whitaker, L. R. (1986). On estimating population characteristics from record-breaking observations, I. Parametric results, *Naval Research Logistics Quarterly*, **33**, 531–543.
20. Samaniego, F. J. and Whitaker, L. R. (1988). On estimating population characteristics from record-breaking observations, II. Nonparametric Results, *Naval Research Logistics Quarterly*, **35**, 221–236.
21. Sibuya, M. and Nishimura, K. (1997). Prediction of record-breakings, *Statistica Sinica*, **7**, 893–906.
22. Yang, T. Y. and Lee, J. C. (2007). Bayesian nearest-neighbor analysis via record value statistics and nonhomogeneous spatial Poisson processes, *Computational Statistics & Data Analysis*, **51**, 4438–4449.